DOCUMENT RESUME

ED 318 746 TM 014 658

AUTHOR Wiser, Berton; Lenke, Joanne M.
TITLE The Stability of Achievement Test Norms over Time.
PUB DATE Apr 87
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987).
AVAILABLE FROM Psychological Corporation, 555 Academic Court, San Antonio, TX 78204 (no charge).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Gains; *Achievement Tests; Comparative Testing; *Elementary School Students; Elementary Secondary Education; *National Norms; *Secondary School Students; Standardized Tests; Statistical Analysis; Statistical Data; Testing Problems; *Test Norms; Test Reliability
IDENTIFIERS *Stanford Achievement Tests

ABSTRACT

The extent to which national performance on an achievement test battery changed over a 4-year period was studied, and the comparative performance of students in school districts taking the test series for the first time was compared with that of students in districts that had used the battery at least once in one or more grades. The seventh edition of the Stanford Achievement Test was used to investigate both issues for 1982 and 1986. After consideration of geographic region, system size, and socioeconomic status, a sample of 116,000 students from kindergarten through grade 12 was selected. Results indicate that national performance did improve in reading, mathematics, and language in the lower grades. Students in lower grades in districts that had used the test series at least once outperformed students in districts that had not used the series before. Use of the 1982 norms to assess current students would overestimate the standing of current students relative to the national population, particularly in mathematics and language. Three tables present the study results. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CD1831
The Stability of Achievement Tes
t Norms over Time.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOANNE M. LENKE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED318746

The Stability of Achievement Test Norms Over Time

Berton Wiser
Joanne M. Lenke

The Psychological Corporation

Paper Presented at the Annual Meeting of the National Council on Measurement in Education. Washington, D.C., April, 1987.

TM014658

BEST COPY AVAILABLE

ABSTRACT

A research study was conducted to investigate the extent to which national performance on an achievement test battery changes over a four-year period, and to examine the comparative performance of students in school districts that were using the particular test series for the first time versus those that had used the battery at least once in one or more grades. The results indicated that national performance did improve in Reading, Mathematics, and Language in the lower school grades. It was also the case that students in school districts that had used the test series at least once outperformed their counterparts in districts that had not used the series before, again in the lower school grades.

The Stability of Achievement Test Norms Over Time

Berton Wiser
Joanne M. Lenke
The Psychological Corporation

When an achievement test series is revised, or when school districts decide to change from one test series to a new edition of a different test series, more often than not, the norm-referenced scores obtained from the "old" test are different from the norm-referenced scores obtained from the "new" test. School districts are then faced with the trauma of explaining observed score changes to their administrations, to the press, and to the public (Lenke and Beck, 1980). The situation has become particularly difficult in recent years, where achievement test norms have become increasingly more "difficult" as test series have been revised. (Actually, more difficult norms reflect the fact that students nationally are performing better now than they did in the 1970's, which reflects the positive impacts of more targeted educational practices.)

Baglin (1981) focused attention on the possibility that achievement test norms could be biased by the disproportionate participation of school districts that are using the previous edition of the achievement test series being standardized. It is also the case that when a school district has been using the same edition of a test series for three, four, five, or more years, scores have a tendency to "creep upwards." That is, test scores tend to improve from year to year, due in part to greater focus on the objectives measured by the test and familiarity on the part of students and teachers with test format and directions for administering.

Therefore, the purpose of the research described in this paper was twofold:

1) To investigate the extent to which national performance on a standardized achievement test battery has changed over a four-year period; and
2) To investigate whether or not students in school districts that have been using a particular achievement test battery outperform students in districts that are administering the battery for the first time.

METHOD

The seventh edition of Stanford Achievement Test (SAT7) was used to investigate both issues. The Stanford Series, which measures achievement in reading, mathematics, language, listening, science, and social studies for students in Kindergarten through Grade 12, consists of ten levels, with two forms at each level except the lowest two. The norms for this series were obtained in the 1981-1982 school year. The research relating to this study was carried out during the spring and fall of 1986; however, only the spring results are reported here, as analysis of the fall data is not yet complete.

The approach taken in the research was to replicate the sampling and analysis procedures used in the original norming of the Stanford Series, but with the addition of a user/non-user stratification category. The sampling and weighting procedures are described below.

Sampling

As was the case in 1982, the 1986 spring sample was designed to give proportionate representation to public and nonpublic schools. Consequently, eighty-nine percent of the national sample was chosen from public school districts. The nonpublic sample was further subdivided proportionately into Roman Catholic schools and all other private schools.

Public school systems were indexed by three demographic variables: geographic region, system size, and socioeconomic status. There were four regions (Northeast, Midwest, Southeast, and West), five system sizes as defined by enrollment per grade (2001 and above, 501-2000, 201-500, 25-200, and less than 25), and five socioeconomic (SES) categories. The SES variable was based on median family income and percent of high school graduates in the community. The five SES categories represent a division of the national population into five groups of equal size.

The application of these stratification variables to a sampling matrix resulted in a 100 cell, 4 x 5 x 5 matrix. Sampling was carried out within each non-empty cell of the matrix so that the proportion of the sample by cell roughly matched the national proportion of public school students within that cell. Within each cell, the chances of a system being selected for participation was proportional to its enrollment.

Nonpublic schools were indexed by two variables: Catholic/Non-Catholic and geographic region. A 2 x 4 sampling matrix was prepared, and, as above, sampling was carried out so that the proportion of the sample from each cell matched the national proportion.

Further, for both the public and nonpublic components, the sample was selected so that approximately thirty-five percent of each had administered the Stanford Series at least once to one or more grades. This percentage was chosen because it represents the estimated market share, on a student basis, of the SAT7 Series.

For each system selected for each cell, two alternate systems were selected from the pool of potential participants. Each alternate was drawn to match (in terms of the stratification variables and its user/non-user status) the original system for which it was an alternate. Such systems were invited automatically whenever a system on the original list declined to participate.

An additional consideration was sampling within systems. Within the largest school systems, of course, not all students could be tested. After each large system agreed to participate in the research, school officials were asked to submit a list of school buildings. A random sample of buildings was then chosen to provide the number of students required (200 per grade). An alternate random sample of buildings was also selected. Again, if one of the original buildings could not participate, an alternate was asked.

All research participants above the Kindergarten level were administered Form S of the Otis-Lennon School Ability Test (OLSAT) in addition to the Stanford. This was done to ensure that the ability level of the 1986 sample would be identical to the ability level of the 1982 sample.

The test administration design is shown in Table 1. The sample sizes shown are based on a desired 2500 students per form per grade for on-level testing and 1500 students per form per grade for lower-level testing. All testing took place over a three-week period, beginning with the third week in April and ending the first week in May, which coincides exactly with the 1982 norming dates.

Table 1. Spring 1986 Stanford Research Program National Sample: Levels and Grades Tested

Stanford Level	Grade	No. of Forms	Number of Students
SESAT 1	K	1	2,500
SESAT 2	K, 1	1	4,000
Primary 1	1, 2	3	12,000
Primary 2	2, 3	2	8,000
Primary 3	3, 4	2	8,000
Intermediate 1	4, 5	2	8,000
Intermediate 2	5, 6, 7	2	13,500
Advanced	7, 8, 9	2	15,000
TASK 1	8, 9, 10, 11, 12	2	25,000
TASK 2	9, 10, 11, 12	2	20,000
			116,000

NOTE: The sample sizes listed above do not include an additional sample of students attending nonpublic schools who were tested to carry out parallel research on a totally nonpublic school sample.

Weighting

Provision was made in the initial planning of the research for statistically weighting the obtained data to achieve a better approximation to national characteristics in terms of the demographic index variables and the desired user/non-user proportions. In addition, since all pupils from Grades 1-12 were administered the OLSAT, it was possible to weight the scores at each grade to match the distributions obtained for the 1982 norming sample.

A separate OLSAT weighting was implemented to match the user and non-user samples in terms of ability. For each grade, level, and form combination, both the non-user and the user groups were weighted so that each of the two samples matched the 1982 and 1986 samples in terms of measured ability.

RESULTS

Table 2 presents a comparison between 1982 and 1986 results for the respective national samples in terms of scaled score means and standard deviations. It is evident from this table that scaled-score gains occurred between 1982 and 1986, the most significant ones showing up in the primary grades in all content areas and in Total Mathematics and Total Language across all grades. (Note that most scaled score differences would be statistically significant because of the very large sample sizes. Therefore, statistical tests of significance were not applied to these data.) Small scaled score reversals occurred only in Total Listening in Grades 6, 7, and 9 and in Social Science in Grade 10.

As far as impact on the national norms is concerned, there is an approximate one-to-one correspondence between a one-point scaled-score increase (or decrease) and a one-point percentile-rank increase (or decrease) at the mean for total scores. Therefore, a difference of seven scaled-score units would represent about seven percentile-rank units at the mean. If one assumes that the seven-point difference exists throughout the score distribution, obviously the percentile-rank difference becomes smaller and smaller as one approaches the tails of the distribution. Based on the data presented in Table 2, it appears that use of the 1982 norms for the Stanford Series overestimates today's students' percentile ranks in many cases.

Table 3 presents the scaled score means and standard deviations for the user and non-user groups, by grade, by content area. As can be seen from this table, users performed as well as or better than non-users in all subject areas through Grade 6. In Grade 7, however, non-users performed better than users in Reading Comprehension; and in Grade 8, non-users performed better than users in Total Mathematics, Total Language, and Science. An unusual pattern emerges in the high school grades: In Grade 9, users performed better than non-users only in Reading Comprehension and Total Listening; in grade 10, only in Total Mathematics; and in Grades 11 and 12, users performed less well than non-users in all subject areas.

Interesting comparisons can be made between the performance of the original 1982 norming sample and that of the 1986 non-user sample, using Tables 2 and 3. These two groups are the same in the sense that 1) the school districts involved had never used this edition of the battery; and 2) the two groups are matched in terms of measured school ability as assessed by OLSAT. Therefore, performance differences between the two groups can be accounted for, to some extent at least, to actual achievement differences. In Total Reading, the performance of the 1982 sample and the 1986 non-user group was the same in Kindergarten and in Grades 4, 5, and 6; however, the non-user group outperformed the 1982 group in all other grades. In Total Mathematics, the non-user group performed better than the 1982 sample in Grades 2 through 12; and in Total Language and in Science, in Grades 3 through 12. In Social Science, the non-user group performed better than the 1982 group in Grades 3 through 7 and in Grades 9, 11, and 12; performance for the two groups was the same in Grades 8 and 10. In Total Listening, on the other hand, there appears to be no consistent pattern across the grades, other than the same or somewhat better performance of the non-user group in Kindergarten through Grade 3 and the same or somewhat poorer performance of this group in Grades 4 through 9.

Table 2. A Comparison of 1982 and 1986 Stanford Scaled Score Means and Standard Deviations

TOTAL READING/
READING (TASK)

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	436.5	48.1	7417	442.9	45.5	11007
1.8	513.1	55.6	13183	522.4	52.5	12422
2.8	574.8	48.5	12421	581.0	46.9	12292
3.8	607.3	46.2	11518	613.5	43.5	11897
4.8	622.5	39.9	13553	623.5	38.8	11781
5.8	640.9	36.7	14501	641.9	36.3	12495
6.8	652.7	35.0	13299	653.6	34.5	11276
7.8*	670.9	46.6	12466	672.2	47.5	12456
8.8*	679.9	46.5	11642	679.5	46.8	12319
9.8*	683.0	45.9	13196	686.2	46.3	8831
10.8	692.0	46.3	10064	694.4	46.2	8671
11.8	703.9	46.3	8860	708.4	45.8	7060
12.8	708.8	46.9	6912	714.7	46.3	5281

TOTAL MATH

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	473.2	47.1	7568	478.5	46.5	11089
1.8	531.8	44.8	13296	547.0	45.8	12387
2.8	558.7	39.8	12666	565.8	39.5	13259
3.8	596.8	42.1	11742	603.3	41.0	11879
4.8	618.5	41.0	12147	623.4	40.0	11724
5.8	643.9	39.6	14435	650.7	40.8	12511
6.8	664.3	40.5	13207	668.5	41.6	11246
7.8	679.0	41.6	12119	682.2	42.3	12178
8.8	698.6	43.8	11428	702.8	46.4	12216
9.8	705.2	43.4	13017	707.2	44.9	8547
10.8	714.0	47.9	10067	715.9	47.64	8604
11.8	725.0	51.6	8653	726.3	51.28	7203
12.8	727.6	52.6	6810	730.4	52.41	5190

*Reading Comprehension

Table 2. A Comparison of 1982 and 1986 Stanford Scaled Score Means and Standard Deviations (continued)

TOTAL LANGUAGE/
TOTAL ENGLISH

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	-	-	-	-	-	-
1.8	-	-	-	-	-	-
2.8	-	-	-	-	-	-
3.8	615.6	46.5	11555	624.8	44.8	11895
4.8	633.7	40.1	13460	638.8	39.4	11758
5.8	651.3	39.8	14361	656.2	39.8	12508
6.8	664.0	38.9	13221	669.2	38.9	11240
7.8	673.4	39.5	12246	678.6	39.2	12184
8.8	687.4	40.9	11424	693.4	40.5	12063
9.8	693.3	39.8	12975	697.9	40.5	8344
10.8	700.4	42.2	9898	704.5	40.4	8537
11.8	709.3	41.5	8444	714.6	40.5	7082
12.8	713.4	42.8	6650	720.3	40.7	5039

TOTAL LISTENING

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	540.7	37.6	7675	544.9	35.2	11376
1.8	563.6	34.3	13255	567.3	37.5	12501
2.8	589.8	34.9	12426	593.0	34.3	13260
3.8	615.9	35.2	11460	618.9	34.2	11840
4.8	632.3	34.7	13392	633.3	33.2	11681
5.8	650.0	35.6	14276	650.0	34.9	12523
6.8	662.0	36.5	13108	661.2	36.1	11213
7.8	671.1	36.1	12090	670.7	35.3	12041
8.8	684.4	35.9	6341	684.9	35.8	6086
9.8	691.4	37.8	4239	689.6	36.6	2776
10.8	-	-	-	-	-	-
11.8	-	-	-	-	-	-
12.8	-	-	-	-	-	-

Table 2. A Comparison of 1982 and 1986 Stanford Scaled Score Means and Standard Deviations (continued).

ENVIRONMENT

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	532.6	36.0	7711	534.4	34.6	11560
1.8	558.4	32.0	13359	559.3	32.9	12240
2.8	581.9	33.7	12694	586.7	32.5	13295
3.8	603.4	37.2	11774	610.9	36.2	11804
4.8	-	-	-	-	-	-
5.8	-	-	-	-	-	-
6.8	-	-	-	-	-	-
7.8	-	-	-	-	-	-
8.8	-	-	-	-	-	-
9.8	-	-	-	-	-	-
10.8	-	-	-	-	-	-
11.8	-	-	-	-	-	-
12.8	-	-	-	-	-	-

SCIENCE

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	-	-	-	-	-	-
1.8	-	-	-	-	-	-
2.8	-	-	-	-	-	-
3.8	606.4	38.7	11794	614.2	37.3	11824
4.8	621.6	40.8	13710	626.6	39.4	11407
5.8	638.6	37.8	14505	642.5	37.2	12559
6.8	651.2	36.6	13263	655.1	36.7	11255
7.8	658.0	36.0	12359	661.0	37.2	12210
8.8	674.2	38.8	11541	677.1	39.2	12103
9.8	679.1	37.9	13234	683.1	39.7	8009
10.8	690.5	42.3	10085	692.7	42.1	8204
11.8	697.0	41.0	8695	700.3	41.6	6587
12.8	698.5	44.6	6901	701.9	43.8	4639

Table 2. A Comparison of 1982 and 1986 Stanford Scaled Score Means and Standard Deviations (continued)

SOCIAL SCIENCE

	1982			1986		
GRADE	$\bar{X}$	SD	N	$\bar{X}$	SD	N
K.8	-	-	-	-	-	-
1.8	-	-	-	-	-	-
2.8	-	-	-	-	-	-
3.8	599.4	38.7	11805	606.6	38.0	11876
4.8	620.9	42.9	13740	623.1	41.3	11410
5.8	637.8	40.7	14512	640.6	39.6	12574
6.8	648.4	39.6	13285	651.5	39.9	11256
7.8	660.0	41.2	12390	661.6	41.3	12229
8.8	682.1	44.3	11516	682.7	44.0	12086
9.8	686.0	42.3	13209	687.3	42.9	7939
10.8	694.1	42.2	10020	692.8	41.6	8374
11.8	704.6	41.3	8616	706.3	41.9	6707
12.8	707.7	45.2	6811	711.1	42.8	4700

Table 3. Scaled Score Means and Standard Deviations for User and Non-user Samples

	Kindergarten		Grade 1		Grade 2		Grade 3	
	Users	Non-Users	Users	Non-Users	Users	Non-Users	Users	Non-Users
TOTAL READING								
$\bar{X}$	455.8	435.9	533.5	518.6	587.5	576.5	616.9	613.0
SD	48.3	44.6	55.9	50.2	46.4	46.6	41.8	43.6
TOTAL MATH								
$\bar{X}$	*		*		569.6	563.4	608.3	601.6
SD					41.6	37.4	41.8	40.7
TOTAL LANGUAGE								
$\bar{X}$							629.4	623.6
SD							38.9	45.3
TOTAL LISTENING								
$\bar{X}$	553.0	540.4	573.6	565.6	598.6	588.9	622.0	619.0
SD	36.3	34.8	32.7	32.9	33.7	34.2	31.9	34.2
ENVIRONMENT								
$\bar{X}$	543.9	529.9	565.8	558.8	588.5	585.5		
SD	36.5	33.9	32.0	32.6	31.1	33.4		
SCIENCE								
$\bar{X}$							620.0	613.1
SD							39.7	37.2
SOCIAL SCIENCE								
$\bar{X}$							610.9	605.6
SD							41.8	37.5

*Not Available

Table 3. Scaled Score Means and Standard Deviations for User and Non-user Samples

	Grade 4		Grade 5		Grade 6		Grade 7	
	Users	Non-Users	Users	Non-Users	Users	Non-Users	Users	Non-Users
TOTAL READING								
$\bar{X}$	626.0	622.3	644.6	640.2	656.6	651.9	669.5**	672.2**
SD	36.3	39.2	34.3	37.4	34.1	34.0	41.2	48.6
TOTAL MATH								
$\bar{X}$	623.0	623.7	651.0	650.4	671.5	666.8	683.2	681.4
SD	40.1	39.8	38.8	42.1	41.9	40.4	39.8	42.2
TOTAL LANGUAGE								
$\bar{X}$	641.0	637.7	657.9	655.1	671.2	667.9	680.2	677.5
SD	39.4	39.2	38.6	40.7	38.6	38.5	36.2	39.8
TOTAL LISTENING								
$\bar{X}$	637.9	631.4	650.6	650.0	664.1	659.4	672.4	670.6
SD	34.0	32.3	34.8	34.7	36.1	35.4	31.6	35.8
ENVIRONMENT								
$\bar{X}$								
SD								
SCIENCE								
$\bar{X}$	628.5	625.9	645.4	640.6	658.7	652.8	663.5	660.5
SD	37.0	40.1	35.0	38.6	36.9	35.9	32.0	38.2
SOCIAL SCIENCE								
$\bar{X}$	625.4	622.8	640.1	641.1	655.1	649.5	665.3	660.9
SD	39.0	41.9	38.0	40.8	38.8	40.0	37.8	41.8

**Reading Comprehension

Table 3. Scaled Score Means and Standard Deviations for User and Non-user Samples

	Grade 8		Grade 9		Grade 10		Grade 11	
	Users	Non-Users	Users	Non-Users	Users	Non-Users	Users	Non-Users
TOTAL READING								
$\bar{X}$	685.2*	676.2*	686.3**	684.5**	694.4	694.4	706.0	709.0
SD	46.4*	46.6*	41.7**	47.2**	43.8	46.9	45.0	46.0
TOTAL MATH								
$\bar{X}$	701.7	703.4	705.6	708.2	717.4	715.3	722.7	727.7
SD	45.1	47.1	42.9	46.0	47.4	47.6	52.0	50.7
TOTAL LANGUAGE								
$\bar{X}$	692.1	694.1	695.0	698.8	702.9	704.9	712.6	715.3
SD	39.4	41.2	33.6	42.4	36.2	41.9	37.6	41.6
TOTAL LISTENING								
$\bar{X}$	686.0	681.5	694.9	689.4				
SD	36.9	32.0	30.2	36.8				
ENVIRONMENT								
$\bar{X}$								
SD								
SCIENCE								
$\bar{X}$	671.7	680.4	684.2	683.8	690.9	693.3	693.7	702.8
SD	35.9	40.7	35.1	41.6	39.2	42.9	37.0	42.7
SOCIAL SCIENCE								
$\bar{X}$	683.3	682.4	686.3	687.9	689.9	693.9	701.9	707.8
SD	43.0	44.6	37.3	43.5	36.5	43.0	41.0	41.8

**Reading Comprehension

Table 3. Scaled Score Means and Standard Deviations for User and Non-user Samples

	Grade 12	
	Users	Non-Users
TOTAL READING		
$\bar{X}$	712.7	715.3
SD	44.9	46.7
TOTAL MATH		
$\bar{X}$	725.7	731.8
SD	53.4	52.0
TOTAL LANGUAGE		
$\bar{X}$	715.6	722.2
SD	36.6	42.3
TOTAL LISTENING		
$\bar{X}$		
SD		
ENVIRONMENT		
$\bar{X}$		
SD		
SCIENCE		
$\bar{X}$	691.4	705.0
SD	38.2	44.6
SOCIAL SCIENCE		
$\bar{X}$	700.2	714.6
SD	40.9	42.4

**Reading Comprehension

DISCUSSION

The results indicate that:

1) The 1986 sample performed better than the 1982 sample as indicated by higher mean scaled scores for the 1986 group when the two groups were matched in terms of measured school ability.

2) Performance improved most in Total Mathematics and Total Language, and least in Total Listening and Social Science. While Total Reading mean scaled scores were higher in 1986 in the primary grades, performance did not dramatically improve in the elementary, intermediate, and high school grades. Though consistent improvement in Science was evident across the grades, the observed gains would have very little impact on the resulting norms.

3) Through Grade 6, users out-performed non-users, even with the ability distributions of those two groups matched.

4) Non-users from the 1986 sample outperformed the 1982 sample, except in Total Listening, where there was no consistent pattern. Again, this was so, even though the ability distributions of the two groups were matched.

There is no evidence that the above results are generalizable to time periods other than the one covered by the research. The results may, in fact, be descriptive only of students from that time frame.

Total Listening and Social Science showed the least scaled score growth between 1982 and 1986. The results are not surprising with regard to Total Listening, as this area is not curriculum based. On the other hand, the lack of change in Social Science, which is curriculum based, is disappointing.

Finally, it is evident that use of the 1982 Stanford norms to assess current students would overestimate the standing of these students relative to the national population. This is particularly the case in Mathematics and Language.

REFERENCE

Baglin, R.F. Self selection bias in national norms. Journal of Educational Measurement, Summer 1981, Volume 18, pp. 97-103.

Lenke, J.M. and Beck, Michael D. The ways and means of test score interpretation. New Directions for Testing and Measurement, 6, 1980, pp. 1-16.